ESIP Lab

SUPPORTING DEVELOPMENT OF EARTH SCIENCES CYBER-INFRASTRUCTURE

Project Summary:

Name of project:

Improve Earth Data Discovery through Deep Query Understanding

Project lead and contact details:

Yongyao Jiang
Ph.D. Candidate in Earth Systems and GeoInformation Sciences
NSF Spatiotemporal Innovation Center, George Mason University
Email: yjiang8@gmu.edu

Project partners (if applicable) and contact details:

Lewis John McGibbney, Ph.D.
Data Scientist II
NASA Jet Propulsion Laboratory/California Institute of Technology
Email: lewis.J.Mcgibbney@jpl.nasa.gov

Justin C. Goldstein, Ph.D.
Research Analyst
National Oceanic & Atmospheric Administration (NOAA) Technology, Planning, and Integration for Observation Program / Riverside Technology, Inc.
Email: justin.goldstein@noaa.gov

Proposed start and end date: 1/1/2018 - 7/31/2018

Budget Requested: $7,000

Budget Summary: To successfully execute the proposed project, we are requesting the full available budget of $7,000 which will be shared equally between project collaborators. The budget will be equally distributed over the project duration. No costs are being directed towards travel or ESIP meeting attendance as these have already been obtained by proposed participating collaborators. Proportional small budgets may be directed towards additional breakout events at ESIP meetings, poster generation, publication, and transition of the project to an appropriate long term home.

This project relates to the following key priority areas (please indicate all that apply):

□ **Earth Science Cyber-infrastructure**

□ **Semantic Technologies**

□ Socioeconomic value of data

□ **Other (explain) - Data/Knowledge Discovery**

Project Outline:

Project description: Discovering Earth science data has been challenging given both the increased quantity and decreased latency of data available and the heterogeneity of the data across a wide variety of domains. One longstanding problem in Earth data discovery is understanding the manner in which one uses existing user queries to interpret the user's search intent. While Google has a "did you mean this…" feature, other search engines are lacking in such technology, especially with regard to the utilization of e.g., fuzzy logic, There are a few existing libraries and APIs available for spatial and temporal parsing. For example, the Google Maps Geocoding API and CLAVIN [1] convert location names into geographic coordinates. Stanford Temporal Tagger (SUTime) [2] can be used for tagging temporal component of a query. Ongoing activity within the Earthdata Search [3], which parses spatial and temporal components from user queries based on CLAVIN and SUTime, is occurring. However, to our knowledge no existing geoinformatics work has tried to parse and tag the non-spatial and temporal components of the query syntax, which usually consists of entities like geophysical variable, satellite name, instrument name, processing level, etc. Understanding the desired objectives behind user queries is difficult because (1) user queries are usually not in full sentences, (2) users tend to use many acronyms in-lieu of full-names, (3) a lack of semantic context exists. Recent progress in deep learning and natural language processing (NLP) algorithms has achieved great performance in query understanding [4] [5] [6]. To fill this gap, we therefore propose to develop a query understanding tool to better interpret users' search intents for Earth data search engines by mining metadata and user query logs. The query understanding tool will be developed through four steps: spatial and temporal parsing, phrase extraction, named entity recognition (NER), and semantic query expansion (Figure 1). Spatial and temporal parsing can be implemented using the existing libraries such as CLAVIN and SUTime. Phrase extraction isolates clauses from free text, which can be conducted by training a language model with a collection of metadata and user search logs. Name entity recognition is used to classify each phrase into a category such as “processing level” or “satellite name”. Likewise, a NER model can be trained with a collection of metadata. Semantic query expansion is used to augment a phrase with its synonyms and acronyms. The query expansion work we have accomplished in the Mining and Utilizing Dataset Relevancy from Oceanographic Datasets to Improve Data Discovery and Access (MUDROD) project [7][8] can be leveraged to accelerate the development of this proposed tool. To demonstrate the query understanding concept, we will utilize our current strong working relationships with NASA JPL’s Physical Oceanography Distributed Archive Center (PO.DAAC).

Project objectives:

This project will focus on the very beginning of the search process - using user queries to identify their search intent, which would in turn improve the search precision, recall, and ranking. The goals of this project include:

- Create an open-source query understanding solution for Earth sciences that fills the aforementioned intent gap in Earth data discovery;
- Provide a proof-of-concept for deep learning and natural language processing in Earth data/knowledge discovery that can be shared with the broader Earth Science Community, including ESIP and interested scientists worldwide through presentations and one or more papers;
- Implement technologies or activities furthering ESIP goals of Discovery, Earth Science Cyber-infrastructure, as well as Semantic Technologies

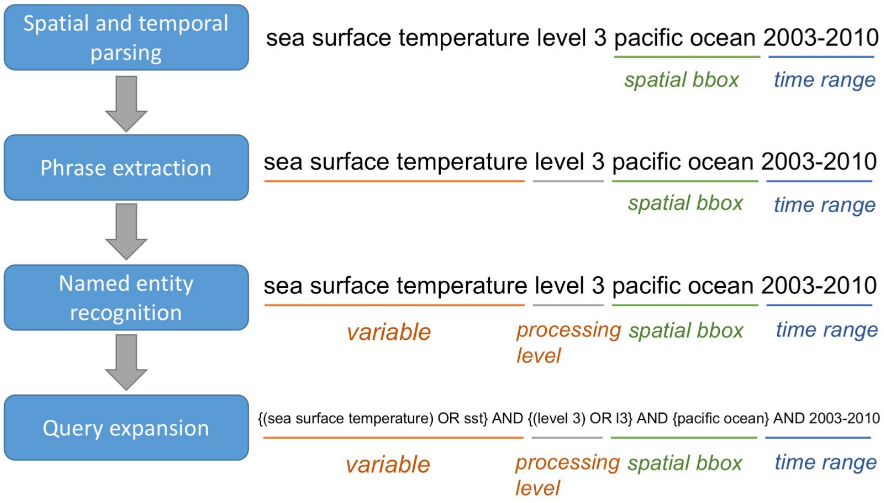
Spatial and temporal parsing
sea surface temperature level 3 pacific ocean 2003-2010
spatial bbox
time range
Phrase extraction
sea surface temperature level 3 pacific ocean 2003-2010
spatial bbox
time range
Named entity recognition
sea surface temperature level 3 pacific ocean 2003-2010
variable
processing level
spatial bbox
time range
Query expansion
{(sea surface temperature) OR sst} AND {(level 3) OR l3} AND {pacific ocean} AND 2003-2010
variable
processing level
spatial bbox
time range

Figure 1. Conceptual example of query understanding

Description of key project steps and timeline:

Task Name	Start	Finish
Project initialization, infrastructure setup, and use case design	11/01/17	11/30/17
Prepare for ESIP winter meeting	11/01/17	01/08/18
Development of the spatial and temporal parsing	11/01/17	01/31/18
Present project plan at ESIP winter meeting	01/08/18	01/10/18
Development of the phrase extraction	02/01/18	03/15/18
Development of the NER function	03/16/18	04/30/18
Development of the query expansion function	05/01/18	06/15/18
Improvement and testing	06/16/18	07/16/18
Develop presentation for ESIP summer meeting	07/17/18	07/31/18

Benefits and advancements of key Earth sciences priority areas:

As a starting point, we will utilize our current strong working relationships with NASA JPL's PO.DAAC and focus on oceanographic data discovery, subsequently moving on to the broader Earth science domains.

Project Partners (as applicable):

Description of project partners and their involvement:

- Lewis John McGibbney: Software framework design, code improvement, and NLP research
- Justin C. Goldstein: Further prototype design, testing, and research
- Matt Austin: Visualization Design

Involvement of ESIP collaboration areas:

- Discovery: Test and provide feedback for the final product
- Semantic Technologies: Provide feed feedback regarding the NER and query expansion results
- Visualisation: Design user driven output to communicate information

Additional Information:

What groups/audiences will be engaged in the project?

- This project would serve as a resource for the NOAA Technology, Planning, and Integration for Observation (TPIO)'s plan to further educate and integrate data management and data science principles within its office, especially its Semantics pilot;.
- This project would provide semantic rich content for decision makers and users of the NOAA Observing System Portfolio for societal benefit;
- The project results can be integrated into MUDROD and Oceanworks (NASA AIST funded projects) [9] to make the the search function of the existing systems more powerful;
- If the project works well with oceanographic data discovery, on which we have already accomplished much with MUDROD (see above), we can reach out to other Earth science data centers represented within ESIP.

How will you judge that project has had impact?

- Positive feedbacks after presentations
- Project results get published in quality journals or conference proceedings

How will you share the knowledge generated by the project?

- Presentation at ESIP/AGU conferences and ESIP cluster/committee monthly calls
- Publication in journals with broad audience such as *International Journal of Geographical Information Science*, and *Earth Science Informatics*

References:

[1] https://github.com/Berico-Technologies/CLAVIN [Accessed 10/10/2017]

[2] https://github.com/wanasit/chrono [Accessed 10/10/2017]

[3]https://wiki.earthdata.nasa.gov/display/EDSC/EDSC+NLP+Overview#?lucidIFH-viewer-ed7bd74b=1 [Accessed 10/10/2017]

[4] Liu, J., Pasupat, P., Wang, Y., Cyphers, S. and Glass, J., 2013, December. Query understanding enhanced by hierarchical parsing structures. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on (pp. 72-77)*. IEEE. doi: 10.1109/BigData.2014.7004310

[5] Demartini, G., Trushkowsky, B., Kraska, T. and Franklin, M.J., 2013, January. CrowdQ: Crowdsourced Query Understanding. In *CIDR*.

[6] AlJadda, K., Korayem, M., Grainger, T. and Russell, C., 2014, October. Crowdsourced query augmentation through semantic discovery of domain-specific jargon. In *Big Data (Big Data), 2014 IEEE International Conference on (pp. 808-815)*. IEEE. doi: 10.1109/BigData.2014.7004310

[7] https://github.com/aist-oceanworks/mudrod [Accessed 10/10/2017]

[8] Jiang, Y., Li, Y., Yang, C., Liu, K., Armstrong, E.M., Huang, T., Moroni, D.F. and Finch, C.J., 2017. A comprehensive methodology for discovering semantic relationships among geospatial vocabularies using oceanographic data discovery as an example. *International Journal of Geographical Information Science*, 31(11), pp.2310-2328. doi:10.1080/13658816.2017.1357819

[9] https://esto.nasa.gov/forum/estf2017/presentations/Huang_A7P6_2017%20ESTF2017.pdf [Accessed 10/10/2017]